
Quality and Performance

Effects of the Premier Hospital Quality Incentive Demonstration on Medicare Patient Mortality and Cost

Andrew M. Ryan

Objective. To evaluate the effects of the Premier Inc. and Centers for Medicare and Medicaid Services Hospital Quality Incentive Demonstration (PHQID), a public quality reporting and pay-for-performance (P4P) program, on Medicare patient mortality, cost, and outlier classification.

Data Sources. The 2000–2006 Medicare inpatient claims, Medicare denominator files, and Medicare Provider of Service files.

Study Design. Panel data econometric methods are applied to a retrospective cohort of 11,232,452 admissions from 6,713,928 patients with principal diagnoses of acute myocardial infarction (AMI), heart failure, pneumonia, or a coronary-artery bypass grafting (CABG) procedure from 3,570 acute care hospitals between 2000 and 2006. Three estimators are used to evaluate the effects of the PHQID on risk-adjusted (RA) mortality, cost, and outlier classification in the presence of unobserved selection, resulting from the PHQID being voluntary: fixed effects (FE), FE estimated in the subset of hospitals eligible for the PHQID, and difference-in-difference-in-differences.

Data Extraction Methods. Data were obtained from CMS.

Principal Findings. This analysis found no evidence that the PHQID had a significant effect on RA 30-day mortality or RA 60-day cost for AMI, heart failure, pneumonia, or CABG and weak evidence that the PHQID increased RA outlier classification for heart failure and pneumonia.

Conclusions. By not reducing mortality or cost growth, this study suggests that the PHQID has made little impact on the value of inpatient care purchased by Medicare.

Key Words. Health care costs, incentives in health care, Medicare

A solution proposed to address the quality and cost problems plaguing health care in the United States is pay-for-performance (P4P), where health care organizations and physicians are compensated not only for what they do, but whether their actions conform to quality standards or result in positive patient outcomes. The promise of P4P has been strong enough to motivate considerable

interest among payers. Recent evidence suggests that the majority of commercial HMOs now use P4P (Rosenthal et al. 2006), most state Medicaid programs use P4P (Kuhmerker and Hartman 2007), and Medicare has launched a number of P4P demonstrations (Centers for Medicare and Medicaid Services [CMS] 2009).

Despite its hope, evidence of the effectiveness of P4P is limited. Several recent reviews have examined the small body of empirical evidence on the effect of explicit financial incentives on the quality of health care and have found little evidence supporting the effectiveness of P4P (Armour et al. 2001; Town et al. 2005; Petersen et al. 2006; Rosenthal and Frank 2006). To the extent that P4P programs have improved performance, improvement has tended to occur for processes of care and not outcomes (e.g., Lindenauer et al. 2007).

To date, P4P has been targeted primarily toward physician practices (Rosenthal et al. 2006). The hospital P4P program that has attracted the most attention as a potential model for a national rollout is the Premier Hospital Quality Incentive Demonstration (PHQID). The PHQID is a collaboration between the CMS and Premier Inc. that began in the fourth quarter of 2003 and continues today. The PHQID pays a 2 percent bonus on Medicare reimbursement rates to hospitals performing in the top decile of performance of a composite quality measure for each clinical condition incentivized in the PHQID (heart failure, acute myocardial infarction [AMI], community-acquired pneumonia, coronary-artery bypass grafting [CABG], and hip and knee replacement) and a 1 percent bonus for hospitals performing in the second highest decile. Penalties for very low performing hospitals were implemented in 2006.

To examine the potential of P4P to improve the value of inpatient care purchased by Medicare, this study will estimate the effect of the PHQID on patient mortality and Medicare cost.

EFFECT OF THE PHQID ON QUALITY: SUMMARY OF EMPIRICAL EVIDENCE

Two published evaluations (Grossbart 2006; Lindenauer et al. 2007) concluded that the PHQID improved quality beyond what would have occurred in its absence while one article found no effect of the PHQID (Glickman et al. 2007). Each of these studies examined the effect of the PHQID among slightly different samples. Because the Lindenauer and colleagues' analysis used the

most complete set of PHQID hospitals and did the most to control for hospital-level confounds, results from this study are the most credible. However, even accepting the Lindenauer and colleagues' results as accurate, doubt remains as to the true impact of the PHQID on patient health. Although seven of the 33 PHQID quality measures are outcomes, the Lindenauer and colleagues' analysis examined only hospital performance on process measures. As noted in other P4P programs, improvement in record keeping or the gaming of measures may be, at least in part, responsible for increased performance on process measures (Petersen et al. 2006). Further, research has shown only a small inverse association between process performance and mortality for AMI, heart failure, and pneumonia among Medicare beneficiaries (Werner and Bradlow 2006; Jha et al. 2007) and also indicates that improvement in process performance has not decreased mortality among Medicare beneficiaries (Ryan et al. 2009, unpublished data). Taken together, this evidence suggests that the PHQID's emphasis on process performance measures may not create incentives that will result in decreases on mortality. The only study that has examined the effect of the PHQID on mortality did so only for AMI, and found no evidence of an effect (Glickman et al. 2007).

POTENTIAL EFFECTS OF THE HQID ON MEDICARE COST

While not previously addressed in the literature, changes in hospital practice in response to the PHQID may also impact Medicare costs. Medicare hospital payment is fixed per admission, based on Diagnosis-Related Groups. However, the PHQID could affect Medicare costs by affecting the number of hospital admissions or hospitals' classification of outliers. (Hospitals are able to receive additional payments for inpatient admissions as a result of particularly costly patients.) To the extent that the PHQID affects admissions (if higher quality care decreases complications and readmissions) or outliers (as a result of greater resource use associated with more intensive treatment or through the gaming of outlier classification,¹ potentially to defray the cost of quality improvement) the PHQID may affect Medicare cost.

METHODS

This analysis estimates the effects of the PHQID on Medicare patient mortality, cost, and outlier classification for AMI, heart failure, pneumonia, and CABG. Although incentivized in the PHQID, hip and knee replacement

are not evaluated because mortality rates are low for these procedures (approximately 0.7 percent). Also, while the financial incentives in the PHQID apply to all patients, not just Medicare beneficiaries, the analysis focuses on Medicare patients because of the availability of mortality and cost data. The econometric approach uses three estimators of the effect of the PHQID to account for unobserved time-varying and time-invariant confounds at the hospital level.

Data

We use several sources of Medicare data from 2000 to 2006: inpatient claims, Denominator files, and Provider of Service files. Inpatient claims are used to identify the principal diagnoses for which beneficiaries are admitted, secondary diagnoses and type of admission for risk adjustment, cost data, and discharge status to exclude transfer patients. The Medicare Denominator File is used to add additional risk adjusters and to determine 30-day mortality. Data from the Medicare Provider of Service file are used to identify hospital structural characteristics. Only short-term, acute care hospitals are included in the analysis. To align the panel with the start of the PHQID, the study period spans the 6-year interval from the fourth quarter of 2000 through the third quarter of 2006. This includes 11,232,452 admissions from 6,713,928 patients with principal diagnoses of AMI, heart failure, pneumonia, or a CABG procedure from the 3,570 acute care hospital entities² included in the analysis.

Dependent Variables

Hospital-level risk-adjusted (RA) 30-day mortality, RA 60-day cost, and RA outlier classification for each incentivized condition are used as the dependent variables. For each admission, 60-day cost is calculated as the sum of each patient's Medicare hospital costs over a 60-day postadmission interval. Costs are attributed to the hospital in which the patient is initially admitted. Classification of outlier status varies substantially over hospitals, with approximately half the hospitals reporting no outliers for AMI, heart failure, and pneumonia over the entire observation period (only around 1 percent of hospitals reported no outliers for CABG). The hospitals that reported no outliers over the observation period are excluded from the outlier analysis.

Risk Adjustment

RA outcomes for each dependent variable are calculated by taking the ratio of the observed outcome to the expected outcome for each hospital, for each

condition, and multiplying this ratio by the grand mean of the outcome for the respective condition. Expected outcomes are estimated by generating predicted values of 30-day mortality and outlier classification from patient-level logit models (and predicted values of 60-day cost from patient-level linear regression models) where the outcome is regressed on age, gender, race, 30 dummy variables for the Elixhauser comorbidities (Elixhauser et al. 1998), type of admission (emergency, urgent, elective), and season of admission. These predicted outcomes are then summed to generate expected outcomes for each hospital for each year.

Modeling the Effects of the PHQID on Mortality, Cost, and Outlier Classification

An important confound in the analysis of the PHQID is the selection effect: because the PHQID is voluntary, hospital participation is a signal of a hospital's interest in improving its clinical quality or reputation. Consequently, quality improvement among hospitals in the PHQID may be the result of these unobserved factors, and not the financial incentives of the PHQID. To address this issue, three approaches are used to estimate the effects of the PHQID on mortality, cost, and outlier classification in the presence of unobserved selection. While it is hypothesized that selection effects will be strongest for mortality, for consistency, each estimator is used for each dependent variable.

First, if unobserved selection can be assumed to remain constant over the observation period, then it can be effectively treated with hospital fixed effects (FE), which control for unobserved time-invariant factors. For condition k at hospital j in year t , the following equation is estimated:

$$RA Outcome_{jkt} = b_1 Year_t + b_2 Year_t \times Z_{jt} + b_3 h_j + \delta PHQID_{jt} + e_{jkt} \quad (1)$$

where $Year$ is a vector of year FE and h is a vector of hospital FE.^{3,4} Z is a vector of hospital characteristics that are hypothesized to have not been affected by the PHQID (ownership, number of beds, medical school affiliation, presence of coronary care unit, presence of inpatient surgical unit, presence of an intensive care unit, presence of open heart surgery facility, and the condition-specific Herfindahl index [a measure of market concentration]). The vector Z varies minimally within hospitals over time and is absorbed by the hospital FE, while the interaction between Z and $Year$ is time-varying and included in the specification. $PHQID$ is equal to 1 if a hospital participated in the PHQID in or after the third quarter of 2003. A negative sign on δ would indicate that the PHQID lowered mortality. Models are estimated separately for each of the PHQID conditions. The mortality and outlier classification models are estimated in levels, while the cost model is estimated in logs to

address the substantial skewness in hospital-level costs. For mortality and outlier classification, semielasticity estimates (measuring the percentage change in the outcome associated with the PHQID) are calculated by dividing the parameter estimate (δ) by the mean of the RA outcome.⁵ Semielasticity estimates for log costs are determined from model coefficients.

The second estimator of the effect of the PHQID relaxes the assumption of constant unobserved selection by estimating the effect of the PHQID in the subsample of only those hospitals that were eligible to participate in the PHQID. As previously noted, participation in the PHQID was voluntary. Of the 421 hospitals asked to participate, 266 (63 percent) chose to do so (Lindenauer et al. 2007). Hospitals' eligibility to participate was based on their subscription to Premier's *Perspective* database, a database used for benchmarking and quality improvement activities, as of March 31, 2003.⁶ While any hospital was eligible to subscribe to the database, subscription is likely indicative of an interest in improving quality. In this specification, time effects (capturing time-varying unobservables) apply only to PHQID-eligible hospitals (for which time effects may be different than for *all* hospitals). Thus, specification 2, which evaluates the effect of the PHQID relative to a set of comparison hospitals that showed an interest in improving their quality, may better account for time-varying confounds than specification 1.⁷ As in specification 1, the mortality and outlier classification models are estimated in levels, while the cost model is estimated in logs.

The third estimator of the effect of the PHQID is the difference-in-difference-in-differences (DDD) estimator. If a hospital's unobserved interest to improve quality applies to multiple domains of its clinical care while the effects of the PHQID are limited to the conditions incentivized under the program, a consistent effect of the PHQID can be estimated from the DDD. The DDD estimator is implemented by modeling the within-hospital difference in outcomes between clinical conditions included and not-included in the PHQID before and after the commencement of the PHQID:

$$\begin{aligned} & \text{std}(\text{RA Outcome HQID Condition}_{jkt}) - \text{std}(\text{RA Outcome Reference Condition}_{jkt}) \\ & = b_1 \text{Year}_t + b_2 \text{Year}_t \times Z_{jt} + b_3 h_j + \delta \text{PHQID}_{jt} + e_{jkt} \end{aligned} \quad (2)$$

where $\text{std}(\cdot)$ denotes the z -score transformation, applied to each RA outcome to normalize outcomes to the same scale across conditions.⁸ A negative sign on δ would indicate that mortality decreased more for PHQID conditions relative to non-PHQID conditions among PHQID hospitals after the implementation

of the PHQID. Because parameter estimates from the DDD models are in units of the standard deviation of RA outcomes, parameter estimates of δ can be interpreted as semielasticities.

Reference Conditions for the DDD

Reference conditions, not incentivized under the PHQID, are used in the DDD specifications to proxy for unobserved time-varying hospital-level interest and effort in improving quality in order to isolate the effect of the financial incentives of the PHQID. Reference condition candidates are chosen from among the Agency for Healthcare Research and Quality (AHRQ) inpatient mortality indicators (AHRQ 2006). The candidate reference conditions are stroke, gastrointestinal hemorrhage, esophageal resection, pancreatic resection, percutaneous transluminal coronary angioplasty (PTCA), carotid endarterectomy (CEA), abdominal aortic aneurysm repair (AAA), and craniotomy.⁹

Four considerations are addressed in the selection of reference conditions for each PHQID condition: (1) the clinical processes of care employed in the treatment of reference conditions must not be subject to spillover effects from the PHQID (resulting in the exclusion of PTCA for AMI patients); (2) a reasonably large number of hospitals must provide treatment for the reference condition in a sufficient volume (resulting in the exclusion of esophageal resection, pancreatic resection, AAA, and craniotomy); (3) mortality rates for reference conditions must be sufficiently large (resulting in the exclusion of non-AMI PTCA and CEA); (4) hospital performance on the conditions must serve as a proxy for overall unobserved hospital interest or effort in improving quality (as determined by a positive, within-hospital, time-varying correlation for 30-day RA mortality between the PHQID and reference conditions). After these exclusions, stroke and gastrointestinal hemorrhage are chosen as the reference conditions for each of the PHQID conditions. Supplemental analysis indicated that inference from the DDD models is not sensitive to the choice of reference conditions (see supporting information Appendix SA2). A detailed description of the methods used to select the reference conditions can be found in supporting information Appendix SA2.

Analysis of Preintervention Trends

To examine whether preintervention trends in RA outcomes were equivalent across the cohorts, supporting the comparison groups as indicators of the counterfactual for PHQID hospitals, simple FE linear trend models are estimated in the preintervention period for mortality, cost, and outlier

classification for each of the incentivized and reference conditions. In these models, RA outcomes are regressed on a time trend and interaction terms between the time trend and dummy variables for hospitals eligible to participate in the PHQID that did not (“eligible hospitals”) and PHQID participating hospitals. A significant interaction term for participating hospitals would indicate nonequivalence of preintervention trends between participating and noneligible hospitals. A significant *difference* in the interaction terms between eligible and participating hospitals would indicate nonequivalence of preintervention trends between these hospitals.

Sensitivity Analysis for Heart Failure and Pneumonia

Circumstances arise in medical care when a patient surviving for an additional 30 days is not a desired health outcome. These circumstances are likely to occur among terminally ill patients who, in the context of the PHQID, are most likely to be heart failure or pneumonia patients.

To evaluate the effect of the PHQID while accounting for possible heterogeneity in performance across terminally ill patients, sensitivity analysis is performed in which the 10 percent of heart failure and pneumonia patients that have the highest predicted probability of mortality, as identified from patient-level logit models, are excluded from the analysis. The previously described estimators of the effect PHQID on heart failure and pneumonia mortality are evaluated among this subset of patients (see supporting information Appendix SA3 for more detail on this sensitivity analysis).

Standard Error Specification

Two sources of heteroskedasticity could arise from these model specifications. First, multiple observations from the same hospitals over time give rise to hospital-level heteroskedasticity. Second, hospital-level RA outcomes vary in their precision as a result of the number of patients in the denominator of the calculation. To treat these two forms of heteroskedasticity, hospital-level cluster-robust standard errors are estimated (Williams 2000) and analytical weights are used based on the cases treated by a hospital, for a given condition, in a given year (Gould 1994). All analysis is performed using *Stata* 10.0 (2007).

RESULTS

Table 1 shows descriptive statistics of the 2006 hospital-level variables used in the analysis. It shows that, relative to hospitals that were not eligible to participate

Table 1: Characteristics of Hospital Cohorts in 2006

| | <i>PHQID Hospitals</i> | <i>PHQID Eligible Nonparticipating Hospitals</i> | <i>Noneligible Hospitals</i> |
|--|------------------------|--|------------------------------|
| <i>N</i> | 256 | 118 | 2,959 |
| Ownership (%) | | | |
| Not-for-profit*** | 85.2 | 83.8 | 59.0 |
| For-profit*** | 2.3 | 4.3 | 20.4 |
| Government run*** | 12.5 | 12.0 | 20.7 |
| Number of beds (%) | | | |
| 1-99*** | 12.1 | 19.5 | 30.6 |
| 100-399 | 57.8 | 61.0 | 53.3 |
| 400+*** | 30.1 | 19.5 | 16.2 |
| Region (%) | | | |
| Northeast* | 12.9 | 11.8 | 16.6 |
| South*** | 53.1 | 47.5 | 41.0 |
| Midwest | 20.7 | 22.9 | 22.4 |
| West*** | 13.3 | 17.8 | 20.0 |
| Medical school affiliation (%)*** | 43.4 | 28.0 | 29.3 |
| Coronary care unit (%)*** | 82.4 | 70.3 | 65.8 |
| Inpatient surgical unit (%) | 96.1 | 94.9 | 94.4 |
| Intensive care unit (%)*** | 96.9 | 98.3 | 90.2 |
| Open-heart surgery facility (%)*** | 52.3 | 33.9 | 30.8 |
| Herfindahl index [median(25th, 75th)] | | | |
| AMI | 0.14(0.08,0.24) | 0.13(0.08, 0.29) | 0.14(0.08, 0.27) |
| Heart failure | 0.09(0.05, 0.16) | 0.09(0.06,0.20) | 0.09(0.05, 0.17) |
| Pneumonia | 0.08(0.04,0.14) | 0.08(0.05, 0.16) | 0.08(0.04,0.13) |
| CABG | 0.31(0.17,0.52) | 0.31(0.19, 0.50) | 0.31(0.18,0.51) |
| Medicare volume [median(25th, 75th)] | | | |
| AMI*** | 99(50,202) | 71(40, 121) | 50(18,117) |
| Heart failure*** | 255(144, 439) | 200(106, 345) | 156(79, 286) |
| Pneumonia*** | 191(138, 292) | 168(100, 229) | 140(83, 222) |
| CABG*** | 26(0, 118) | 0(0,56) | 0(0,41) |
| RA 30-day mortality rate [median(25th, 75th)] | | | |
| AMI*** | 14.4(11.6, 19.3) | 16.9(12.4, 21.8) | 16.9(12.1, 24.8) |
| Heart failure** | 9.0(7.5, 10.8) | 9.7(7.2, 11.1) | 9.6(7.6, 12.1) |
| Pneumonia | 11.0(8.9, 14.1) | 11.1(9.0, 13.6) | 11.3(8.9, 14.3) |
| CABG** | 3.7(2.4, 5.4) | 2.8(1.8, 5.2) | 3.9(2.3, 5.7) |
| Stroke | 17.7(14.7, 20.9) | 17.9(14.3, 21.9) | 17.4(13.4, 21.8) |
| Gastrointestinal hemorrhage | 6.5(5.1, 8.6) | 6.2(4.4, 9.1) | 6.7(4.3, 9.6) |
| RA 60-day cost in thousands of U.S. dollars [median(25th, 75th)] | | | |
| AMI*** | 25.1(20.4, 34.1) | 24.6(20.5, 32.8) | 27.1(20.9, 37.9) |
| Heart failure*** | 13.4(11.3, 16.4) | 12.4(10.8, 13.8) | 13.1(11.0, 16.1) |
| Pneumonia | 9.8(8.8, 11.6) | 9.7(8.6, 11.2) | 10.1(8.7, 12.3) |
| CABG* | 37.3(33.6, 44.8) | 34.7(31.1, 40.1) | 37.8(33.1, 45.7) |
| Stroke | 16.1(13.1, 19.4) | 15.2(12.6, 18.2) | 15.9(12.4, 20.1) |
| Gastrointestinal hemorrhage* | 9.5(8.1, 11.1) | 9.1(7.8, 10.4) | 9.6(8.1, 11.8) |

continued

Table 1. *Continued*

| | <i>PHQID Hospitals</i> | <i>PHQID Eligible Nonparticipating Hospitals</i> | <i>Noneligible Hospitals</i> |
|---|----------------------------|--|----------------------------------|
| RA outlier classification rate [median(25th, 75th)] | | | |
| AMI*** | 2.5(0, 4.7) | 0.1(0, 3.2) | 0(0, 3.8) |
| Heart failure*** | 1.3(0.3, 2.6) | 0.1(0, 1.7) | 0.1(0, 1.7) |
| Pneumonia*** | 1.2(0.4, 2.4) | 0.1(0, 2.0) | 0.1(0, 1.8) |
| CABG | 10.0(6.4, 14.1) | 10.0(3.9, 15.6) | 9.2(5.2, 15.4) |
| Stroke*** | 1.3(0, 2.7) | 0(0, 1.9) | 0(0, 2.0) |
| Gastrointestinal hemorrhage*** | 1.3(0, 2.4) | 0.1(0, 1.9) | 0(0, 2.1) |

*** $p < .01$. ** $p < .05$. * $p < .10$.

Data on hospital characteristics comes from the 2006 Medicare Provider of Service file that is drawn from surveys that may be up to 3 years old.

For ownership, bed size, region, medical school affiliation, coronary care unit, inpatient surgical unit, intensive care unit, and open-heart surgery facility, Wald tests on the equality of proportions across cohorts are performed.

For Herfindahl indices, Medicare volume, RA 30-day mortality, RA 60-day cost, and RA outlier status, nonparametric K -sample tests on the equality of medians across cohorts are performed.

AMI, acute myocardial infarction; CABG, coronary-artery bypass grafting; PHQID, Premier Hospital Quality Incentive Demonstration; RA, risk adjusted.

in the PHQID, a higher proportion of PHQID hospitals are not for profit, have 400 or more beds, have a medical school affiliation, and have an open-heart surgery facility. It also shows that PHQID hospitals have higher Medicare volume for the incentivized conditions, lower mortality for AMI and heart failure, and higher outlier classification for AMI, heart failure, and pneumonia. Eligible hospitals have a similar proportion of not-for-profit ownership as PHQID hospitals but tend to be smaller, less likely to have a medical school affiliation, and less likely to have an open-heart surgery facility. Eligible hospitals have Medicare volume between PHQID and noneligible hospitals and have mortality and outlier classification rates similar to noneligible hospitals. Table 1 supports the idea that PHQID hospitals are substantially different than the other hospital cohorts, particularly noneligible hospitals, both with regard to structural characteristics and the study outcomes.

Figure 1 shows the RA 30-day mortality rates from 2001 to 2006 for the incentivized and reference conditions for each cohort of hospitals. In this figure, the year 2001 spans the fourth quarter of 2000 through the third quarter of 2001. This convention is followed for each year. Mortality for the reference conditions is not displayed for eligible hospitals because the reference conditions are not examined in this cohort.

Figure 1: Risk-Adjusted 30-Day Mortality for PHQID, PHQID Eligible, and Noneligible Hospitals from 2001 to 2006

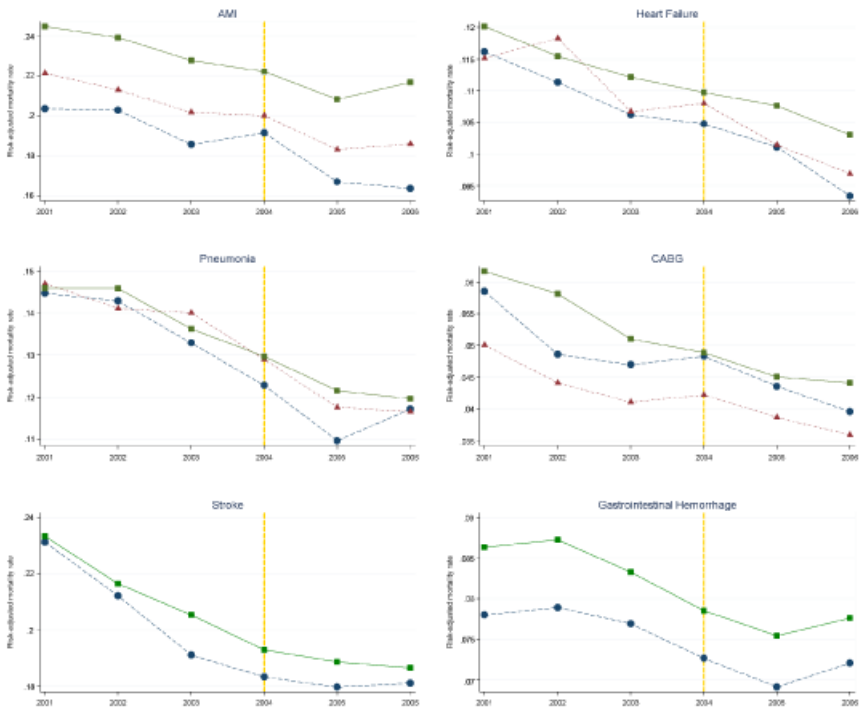


Figure 1 shows that while RA mortality decreased for the PHQID hospitals for each of the incentivized conditions over the study period, similar decreases are observed for eligible and noneligible hospitals. Figure 1 also shows that mortality for stroke and gastrointestinal hemorrhage, which were not incentivized, decreased substantially for PHQID hospitals from 2001 to 2006. For each condition shown in Figure 1, a substantial discontinuity or change in slope is not apparent for PHQID hospitals at the commencement of the intervention.

Results from the FE linear trend models (not shown) indicate that, on an annual basis in the preintervention period, RA 30-day mortality decreased significantly for each of the incentivized and reference conditions, RA 60-day costs *decreased* significantly for CABG and stroke but increased significantly for the other conditions, and RA outlier classification decreased significantly for

each of the incentivized and reference conditions. However, differences in preintervention trends in RA outcomes between PHQID hospitals and comparison hospitals (eligible and noneligible) tended to be small and not statistically different at $p < .05$, supporting the comparison hospitals as indicators of the counterfactual for PHQID participants (see endnote 7 for a description of RA mortality trends over the entire study period).

Table 2 shows the results of the three estimators of the effect of the PHQID on RA 30-day mortality. The first row shows the results from the FE specification estimated among the entire sample of hospitals (Equation 1), the second row shows the results from the FE specification estimated among the subset of hospitals eligible to participate in the PHQID, and rows 3 and 4 show the results from the DDD estimator estimated among all hospitals (Equation 2). For the DDD results, the separate rows correspond to the use of different reference conditions. Table 2 shows that, while the parameter estimates tend to be negative, particularly for AMI, heart failure, and pneumonia, the effect of the PHQID on mortality is small and nonsignificant for each of the estimators. Sensitivity analysis (not shown) that excluded the sickest 10 percent of heart failure and pneumonia patients yielded very similar results and identical parameter inference to models estimated among all patients.

Table 3 shows that the PHQID did not affect RA 60-day Medicare costs for any of the incentivized conditions for the FE models estimated in the entire sample and the PHQID-eligible subsample. However, it shows that in the DDD models where stroke is used as a reference condition, the PHQID is associated with significantly higher costs for heart failure, pneumonia, and CABG. This finding is a result of the relative reduction in 60-day costs for stroke for PHQID hospitals after the commencement of the PHQID, not the result of relative increases in costs for heart failure, pneumonia, and CABG for PHQID hospitals (see Figure C1 in supporting information Appendix SA4).

Table 4 shows estimates of the effect of the PHQID on hospital classification of outliers. It shows no evidence that the PHQID affected the classification of outliers for AMI or CABG. However, in the FE models estimated among all of the hospitals, the PHQID is significantly ($p < .01$) associated with a 41.8 and 24.9 percent increase in outlier classification for heart failure and pneumonia, respectively. A smaller, nonsignificant increase is also observed for heart failure and pneumonia in the FE models estimated among eligible hospitals. In the DDD specifications, the PHQID is not associated with significant increases in outlier classification for heart failure or pneumonia for either of the reference conditions.

Table 2: Estimates of the Effect of the Premier Hospital Quality Demonstration on Risk-Adjusted 30-Day Mortality

| Description | AMI Mortality | | | Heart Failure Mortality | | | Pneumonia Mortality | | | CABG Mortality | | |
|--|-------------------------------------|-------------------|----------------|---------------------------|-------------------|----------------|---------------------------|-------------------|----------------|-------------------------|------------------|----------------|
| | PHQID Effect | N | R ² | PHQID Effect | n | R ² | PHQID Effect | n | R ² | PHQID Effect | n | R ² |
| Hospital fixed effects | -0.002 [-0.9%] (0.002) | 19,729 {3,438} | 0.070 | -0.000 [-0.3%] (0.001) | 20,366 {3,524} | 0.081 | -0.001 [-0.8%] (0.002) | 20,436 {3,538} | 0.142 | 0.002 [4.8%] (0.002) | 6,460 {1,152} | 0.087 |
| Hospital fixed effects in PHQID-eligible subsample | -0.003 [-1.3%] (0.004) | 2,229 {374} | 0.170 | 0.002 [1.9%] (0.002) | 2,243 {375} | 0.166 | 0.001 [0.9%] (0.003) | 2,246 {376} | 0.253 | 0.002 [5.0%] (0.003) | 1,042 {178} | 0.130 |
| | <i>std(AMI Mortality)</i> | | | | | | | | | | | |
| DDD | <i>std(Heart Failure Mortality)</i> | | | | | | | | | | | |
| Stroke | -0.010 [-1.0%] (0.012) | 19,673 {3,427} | 0.026 | -0.003 [-0.3%] (0.011) | 20,084 {3,480} | 0.031 | -0.001 [-0.1%] (0.012) | 20,091 {3,483} | 0.011 | 0.024 [2.4%] (0.028) | 6,449 {1,150} | 0.014 |
| Gastrointestinal hemorrhage | -0.011 [-1.1%] (0.008) | 19,681 {3,429} | 0.031 | -0.006 [-0.6%] (0.008) | 20,175 {3,497} | 0.019 | -0.008 [-0.8%] (0.009) | 20,190 {3,501} | 0.062 | 0.036 [3.6%] (0.027) | 6,445 {1,151} | 0.059 |

*** $p < .01$. ** $p < .05$. * $p < .10$.

Percentage change (semielasticity) in risk-adjusted mortality is displayed in brackets [].

Heteroskedasticity-consistent standard errors are displayed in parentheses ().

Unique numbers of hospital units are displayed in { } under model n .

PHQID and reference conditions are z -transformed in the DDD models.

AMI, acute myocardial infarction; DDD, difference-in-difference-in-differences; PHQID, premier hospital quality incentive demonstration.

Table 3: Estimates of the Effect of the Premier Hospital Quality Demonstration on Risk-Adjusted 60-Day Cost

| Description | ln(AMI Cost) | | | ln(Heart Failure Cost) | | | ln(Pneumonia Cost) | | | ln(CABG Cost) | | |
|--|--|-------------------|----------------|---------------------------|-------------------|----------------|---------------------------|-------------------|----------------|---------------------------|------------------|----------------|
| | PHQID Effect | n | R ² | PHQID Effect | N | R ² | PHQID Effect | n | R ² | PHQID Effect | n | R ² |
| Hospital fixed effects | -0.006 [-0.6%] (0.010) | 19,729 {3,438} | 0.067 | 0.008 [0.8%] (0.009) | 20,364 {3,524} | 0.371 | -0.006 [-0.6%] (0.007) | 20,434 {3,537} | 0.101 | 0.016 [1.6%] (0.014) | 6,460 {1,152} | 0.087 |
| Hospital fixed effects in PHQID-eligible subsample | -0.009 [-0.9%] (0.018) | 2,229 {374} | 0.121 | 0.009 [0.9%] (0.015) | 2,242 {375} | 0.478 | -0.003 [-0.3%] (0.013) | 2,246 {376} | 0.091 | -0.008 [-0.8%] (0.023) | 1,042 {178} | 0.197 |
| DDD | <i>std(AMI Cost)</i> <i>std(Heart Failure Cost)</i> <i>std(Pneumonia Cost)</i> <i>std(CABG Cost)</i> | | | | | | | | | | | |
| Stroke | 0.008 [0.8%] (0.022) | 19,673 {3,427} | 0.037 | 0.057 [5.7%]** (0.024) | 20,084 {3,480} | 0.113 | 0.045 [4.5%]** (0.020) | 20,091 {3,483} | 0.033 | 0.088 [8.8%]** (0.043) | 6,449 {1,150} | 0.029 |
| Gastrointestinal hemorrhage | -0.025 [-2.5%] (0.017) | 19,681 {3,429} | 0.007 | 0.013 [1.3%] (0.017) | 20,175 {3,497} | 0.088 | 0.002 [0.2%] (0.013) | 20,190 {3,501} | 0.010 | 0.058 [5.8%] (0.043) | 6,445 {1,151} | 0.022 |

*** $p < .01$. ** $p < .05$. * $p < .10$.

Percentage change (semielasticity) in risk-adjusted cost is displayed in brackets [].

Heteroskedasticity-consistent standard errors are displayed in parentheses ().

Unique numbers of hospital units are displayed in { } under model n .

PHQID and reference conditions are z-transformed in the DDD models.

AMI, acute myocardial infarction; CABG, coronary-artery bypass grafting; DDD, difference-in-difference-in-differences; PHQID, premier hospital quality incentive demonstration.

Table 4: Estimates of the Effect of the Premier Hospital Quality Demonstration on Risk-Adjusted Outlier Classification

| Description | AMI Outlier | | Heart Failure Outlier | | Pneumonia Outlier | | CABG Outlier | | | | |
|--|--------------------------|-------------------|-----------------------------------|-----------------------------|-------------------------------|----------------|-----------------------------|-------------------|----------------|----------------------------|------------------|
| | PHQID Effect | n | R ² | PHQID Effect | n | R ² | PHQID Effect | N | R ² | | |
| Hospital fixed effects | 0.004 [14.2%] (0.004) | 14,540 {2,479} | 0.191 | 0.006 [41.8%]*** (0.002) | 16311 {2,780} | 0.149 | 0.004 [24.9%]*** (0.002) | 17,349 {2,957} | 0.146 | 0.016 [11.3%] (0.010) | 6,379 {1,136} |
| Hospital fixed effects in PHQID-eligible subsample | 0.000 [-0.7%] (0.005) | 1,948 {325} | 0.267 | 0.003 [17.3%] (0.004) | 2,052 {343} | 0.184 | 0.003 [19.0%] (0.003) | 2,099 {351} | 0.235 | -0.013 [-10.2%] (0.017) | 1,024 {175} |
| | <i>std(AMI Outlier)</i> | | <i>std(Heart Failure Outlier)</i> | | <i>std(Pneumonia Outlier)</i> | | <i>std(CABG Outlier)</i> | | | | |
| DDD | | | | | | | | | | | |
| Stroke | 0.013 [1.3%] (0.035) | 14,519 {2,476} | 0.087 | 0.001 [0.1%] (0.011) | 16,260 {2,772} | 0.097 | 0.003 [0.3%] (0.009) | 17,289 {2,951} | 0.063 | 0.048 [4.8%] (.037) | 6,368 {1,134} |
| Gastrointestinal hemorrhage | 0.022 [2.2%] (0.038) | 14,516 {2,476} | 0.163 | 0.015 [1.5%] (0.008) | 16,267 {2,775} | 0.025 | 0.010 [1.0%] (0.006) | 17,307 {2,955} | 0.025 | 0.061 [6.1%] (0.039) | 6,364 {1,135} |

****p* < .01. ***p* < .05. **p* < .10.

Percentage change (semielasticity) in risk-adjusted outlier payment is displayed in brackets [].

Heteroskedasticity-consistent standard errors are displayed in parentheses ().

Unique numbers of hospital units are displayed in { } under model *n*.

PHQID and reference conditions are *z*-transformed in the DDD models.

AMI, acute myocardial infarction; CABG, coronary-artery bypass grafting; DDD, difference-in-difference-in-differences; PHQID, premier hospital quality incentive demonstration.

DISCUSSION

Using three estimators to account for selection effects and other time-varying and time-invariant confounds, this analysis found no evidence that the PHQID had a significant effect on RA 30-day mortality for AMI, heart failure, pneumonia, or CABG. Also, while the DDD estimator using stroke as the reference condition indicates that the PHQID increased RA 60-day Medicare inpatient cost for heart failure, pneumonia, and CABG, this finding is a result of a reduction in the growth of stroke costs for PHQID hospitals: because it is extremely unlikely that the PHQID caused a reduction in the growth of stroke costs, evidence that the PHQID had a causal impact on 60-day cost is very weak. Semielasticity estimates of the effect of the PHQID on mortality and cost are also very small, typically <2 percent, indicating that large standard errors are not chiefly responsible for the null results. Further, because (1) no evidence of an effect of PHQID effect on mortality is observed in specification 1, which likely underestimates the effect of the PHQID by under-accounting for PHQID hospitals' unobserved interest in improving quality and because (2) the effect of the PHQID in specification 2 is likely biased *away* from the null because PHQID hospitals opted to participate in P4P and other premier hospitals opted not to participate, the effect of the PHQID on the mortality is likely *overestimated* in the analysis. This strengthens the inference that the PHQID did not reduce mortality.

The analysis also provides limited evidence that the PHQID increased RA outlier classification for heart failure and pneumonia. While the effect of the PHQID on outlier classification is large and significant for heart failure and pneumonia in the FE models estimated among all hospitals, the other estimators of the effect of the PHQID on outlier classification are nonsignificant.

Taken together, the analysis suggests that the PHQID has had no causal effect on mortality or Medicare cost for AMI, heart failure, pneumonia and CABG and that evidence of the causal effect of the PHQID on outlier classification for heart failure and pneumonia is weak. Even if the PHQID had a causal effect on outlier classification for heart failure and pneumonia, this effect was not large enough to be reflected in the effect of the PHQID on 60-day cost. Although RA mortality for AMI and heart failure is significantly lower for PHQID hospitals relative to eligible and noneligible hospitals in 2006, after the PHQID was implemented, this does not appear to be a result of the financial incentives of the PHQID.

For AMI, this study confirms the results from Glickman and colleagues who found that the PHQID did not reduce mortality. For heart failure,

pneumonia, and CABG, this study is the first to show that the PHQID had no impact on mortality. This study is also the first to show that the PHQID had no impact on Medicare cost for any of the incentivized conditions examined.

This study has several relevant limitations. First, the study relied on 30-day mortality as the only health outcome of the PHQID and was performed using data from only Medicare patients. Analysis of other health outcomes and analysis using data from non-Medicare patients may have resulted in different conclusions. Second, it could be argued that 30-day mortality and 60-day cost are inappropriate indicators of performance for hospitals in the PHQID, given that the effects of the PHQID may be longer term and not observable within these time intervals. However, adherence with most of the performance measures used in the PHQID would be expected to have a short-term impact on mortality for the incentivized conditions (Niederman et al. 2001; Hunt et al. 2005; Anderson et al. 2008). For example, in the case of AMI, adherence to six out of eight process measures (aspirin at arrival, aspirin prescribed at discharge, angiotension converting enzyme inhibitor for left ventricular systolic dysfunction, β -blocker at arrival, β -blocker prescribed at discharge, thrombolytic agent received within 30 minutes of hospital arrival, percutaneous coronary intervention received within 120 minutes of hospital arrival) would be expected to have a short-term impact on mortality, while two measures (aspirin prescribed at discharge and adult smoking cessation advice/counseling) likely would not. Inpatient mortality is also a quality measure for AMI in the PHQID. Because mortality is frequently preceded by costly care (Lubitz and Riley 1993), it is expected that cost outcomes would similarly be affected by these process measures in the short term. As a result, it is reasonable to evaluate 30-day mortality and 60-day cost as outcome measures that are potentially impacted by the PHQID.

Third, the analysis used a relatively limited set of patient-level demographic and comorbidity data to derive hospital-level RA mortality rates. However, to the extent that unobserved patient severity was constant within hospitals over time, it would be accounted for in the hospital FE. Fourth, under a multitasking scenario, in which hospital effort is allocated toward the incentivized conditions and away from nonincentivized conditions, or a spillover scenario, in which improvements in quality for the conditions incentivized under the PHQID would carry over to nonincentivized conditions, the assumption underlying the DDD, that the PHQID had no effect on nonincentivized conditions, would be invalid. Figure 1 indicates that mortality rates for PHQID participants follow similar trends to noneligible hospitals immediately before and after the PHQID began for the nonincentivized

conditions (stroke and gastrointestinal hemorrhage). This could mean that neither multitasking nor spillover effects occurred or that both multitasking and spillover effects occurred. While it is difficult to determine if the DDD is biased by multitasking or spillover effects, the estimator adds to the robustness of the results by supporting the conclusions from the other models while using an alternative approach toward addressing time-varying unobservables.

Finally, the relatively small number of hospitals that participated in the PHQID and that were eligible to participate limited the power of the analysis to detect small effects of the PHQID on the outcomes investigated. For example, low power likely resulted in a nonsignificant estimate of the effect of the PHQID on outlier classification in the FE models estimated among the subset of PHQID eligible hospitals for heart failure and pneumonia, despite the relatively large magnitude of this effect. However, in the case of mortality and cost, the directionality of the effect of the PHQID is inconsistent across the estimators, suggesting that a more high-powered study would not conclude that the PHQID reduced mortality or cost growth.

CONCLUSION

The use of P4P programs by public and private payers has been rapidly expanding as a strategy to improve quality of care without further increasing cost growth. The PHQID is frequently cited as a model for other P4P programs. However, findings from this study indicate that, by not reducing mortality or cost growth, the PHQID has made little impact on the value of inpatient care purchased by Medicare. Despite these findings, it is premature to brand the P4P enterprise a failure. While empirical evidence is limited, conceptually, the design of P4P programs, including performance measures, payout rules, and the magnitude of rewards, likely affect providers' responsiveness to the financial incentives (IOM 2006). By using primarily process measures, the PHQID may not have been sufficiently targeted to decrease mortality. Also, the magnitude of the financial incentives in the PHQID may have been insufficient to defray the high cost of outcome improvement (Romley and Goldman 2008).

As P4P for inpatient care is rolled out nationally as part of Value-Based Purchasing (U.S. Congress 2005), Medicare should test a variety of P4P designs regionally in order to refine incentives for quality improvement and cost reduction. Designing a P4P program to improve health care quality in a cost-efficient manner in the context of the existing complexity and competing

incentives in the U.S. health care system is an exceedingly difficult undertaking, but one worthy of continued thought and energy.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This work has been supported by a training grant from Agency for Health Care Research and Quality (grant number 05 T32 HS000062-14) and by the Jewish Healthcare Foundation under the grant “Achieving System-wide Quality Improvements—A collaboration of the Jewish Healthcare Foundation and Schneider Institutes for Health Policy.” I would like to thank Stanley Wallack, Christopher Tompkins, Deborah Garnick, Christopher Baum, and Mandy Smith for helpful comments on this paper.

Disclosures: None.

Disclaimers: None.

NOTES

1. See the following for information on hospital gaming of outlier payments: <http://www.hhs.gov/asl/testify/t030311.html>
2. Unique hospital entities are identified by their Medicare provider number.
3. Hospital FE are assigned based on hospitals’ Inpatient Prospective Payment System identification number. If this number changes for a hospital, as a result of a merger, for example, a separate FE is assigned.
4. Hausman tests of the consistency of random effects are rejected in some models, but not all. To ensure comparability across models, FE are specified in each model.
5. Approximately 25 percent of the remaining observations have values of RA outlier classification of 0 for AMI, heart failure, and pneumonia, giving rise to the problem of censored dependent variables and biased parameter estimates. To address this issue, random-effects Tobit models are estimated for RA outlier classification. Results from these models (not shown) are very similar to those from the FE models.
6. See the following for details <http://www.premierinc.com/quality-safety/tools-services/p4p/hqi/faqs-year1-3.jsp#eligible>
7. To examine potential differences in time trends between specifications 1 and 2, the models were reestimated with a linear time trend substituted for the time dummies. For each incentivized condition, the time trend is more negative when estimated among hospitals eligible for the PHQID, relative to the cohort of all hospitals (although it is not significant in specification #2 for AMI). This indicates that hospitals eligible to participate in the PHQID were on a trajectory to decrease mortality more than the cohort of all hospitals, supporting the estimation of specification 2.

8. Results from DDD models estimated using untransformed RA outcomes and logged RA outcomes are virtually identical to those reported for the z -transformed outcomes.
9. Hip fracture mortality is an AHRQ mortality indicator that is excluded because of its close correspondence with hip replacement mortality, a PHQID condition.

REFERENCES

- Agency for Health Care Research and Quality (AHRQ). 2006. "Inpatient Quality Indicators Fact Sheet" [accessed on September 20, 2007]. Available at <http://www.qualityindicators.ahrq.gov/downloads/iqi/2006-Feb-InpatientQualityIndicators.pdf>
- Anderson, J. L., C. D. Adams, and E. A. Antman. 2008. "ACC/AHA 2007 Guidelines for the Management of Patients with Unstable Angina/Non-ST-Elevation Myocardial Infarction: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing committee to revise the 2002 guidelines for the management of patients with unstable angina/non-ST-elevation myocardial infarction) (vol. 50, p. 1, 2007)." *Journal of the American College of Cardiology* 51 (9): 974–84.
- Armour, B. S., M. M. Pitts, R. Maclean, C. Cangialose, M. Kishel, H. Imai, and J. Etchason. 2001. "The Effect of Explicit Financial Incentives on Physician Behavior." *Archives of Internal Medicine* 161 (10): 1261–66.
- Centers for Medicare and Medicaid Services (CMS). 2009. "Premier Hospital Quality Incentive Demonstration" [accessed on January 19, 2007]. Available at http://www.cms.hhs.gov/HospitalQualityInits/35_HospitalPremier.asp
- Elixhauser, A., C. Steiner, D. R. Harris, and R. N. Coffey. 1998. "Comorbidity Measures for Use with Administrative Data." *Medical Care* 36 (1): 8–27.
- Glickman, S. W., F. S. Ou, E. R. DeLong, M. T. Roe, B. L. Lytle, J. Mulgund, J. S. Rumsfeld, W. B. Gibler, E. M. Ohman, K. A. Schulman, and E. D. Peterson. 2007. "Pay for Performance, Quality of Care, and Outcomes in Acute Myocardial Infarction." *Journal of the American Medical Association* 297 (21): 2373–80.
- Gould, W. 1994. "Clarification on Analytic Weights with Linear Regression." *Stata Technical Bulletin* 20: 2–3.
- Grossbart, S. R. 2006. "What's the Return? Assessing the Effect of "Pay-for-Performance" Initiatives on the Quality of Care Delivery." *Medical Care Research and Review* 63 (1): 29S–48S.
- Hunt. 2006. "ACC/AHA 2005 Guideline Update for the Diagnosis and Management of Chronic Heart Failure in the Adult: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing committee to update the 2001 guidelines for the evaluation and management of heart failure) (vol. 112, p. e154, 2005)." *Circulation* 113 (13): E684–85.

- Institute of Medicine (IOM) 2006. *Rewarding Provider Performance: Aligning Incentives in Medicare*. Washington, DC: National Academy Press.
- Jha, A. K., E. J. Orav, Z. H. Li, and A. M. Epstein. 2007. "The Inverse Relationship between Mortality Rates and Performance in the Hospital Quality Alliance Measures." *Health Affairs* 26: 1104–10.
- Kuhmerker, K., and T. Hartman. 2007. *Pay-for-Performance in State Medicaid Programs: A Survey of State Medicaid Directors and Programs*. New York: The Commonwealth Fund.
- Lindenauer, P. K., D. Remus, S. Roman, M. B. Rothberg, E. M. Benjamin, A. Ma, and D. W. Bratzler. 2007. "Public Reporting and Pay for Performance in Hospital Quality Improvement." *New England Journal of Medicine* 356 (5): 486–96.
- Lubitz, J. D., and G. F. Riley. 1993. "Trends in Medicare Payments in the Last Year of Life." *New England Journal of Medicine* 328 (15): 1092–6.
- Niederman, M. S., L. A. Mandell, A. Anzueto, J. B. Bass, W. A. Broughton, G. D. Campbell, N. Dean, T. File, M. J. Fine, P. A. Gross, F. Martinez, T. J. Marrie, J. F. Plouffe, J. Ramirez, G. A. Sarosi, A. Torres, R. Wilson, and V. L. Yu. 2001. "Guidelines for the Management of Adults with Community-Acquired Pneumonia—Diagnosis, Assessment of Severity, Antimicrobial Therapy, and Prevention." *American Journal of Respiratory and Critical Care Medicine* 163 (7): 1730–54.
- Petersen, L. W. L. D., T. Urech, C. Daw, and S. Sookanan. 2006. "Does Pay-for-Performance Improve the Quality of Health Care?" *Annals of Internal Medicine* 145: 265–72.
- Romley, J. A., and D. Goldman. 2008. *How Costly Is Hospital Quality? A Revealed-Preference Approach*. NBER Working Paper No. 13730.
- Rosenthal, M. B., and R. G. Frank. 2006. "What Is the Empirical Basis for Paying for Quality in Health Care?" *Medical Care Research and Review* 63 (2): 135–57.
- Rosenthal, M. B., B. E. Landon, S. T. Normand, R. G. Frank, and A. M. Epstein. 2006. "Pay for Performance in Commercial HMOs." *New England Journal of Medicine* 355 (18): 1895–902.
- StataCorp. 2007. *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP.
- Town, R., R. Kane, P. Johnson, and M. Butler. 2005. "Economic Incentives and Physicians' Delivery of Preventive Care—A Systematic Review." *American Journal of Preventive Medicine* 28 (2): 234–40.
- U.S. Congress. 2005. "House Report 109-362 Deficit Reduction Act of 2005" [accessed on January 29, 2008]. Available at [http://thomas.loc.gov/cgi-bin/cpquery/R?cp109:FLD010:@1\(hr362\)](http://thomas.loc.gov/cgi-bin/cpquery/R?cp109:FLD010:@1(hr362))
- Werner, R. M., and E. T. Bradlow. 2006. "Relationship between Medicare's Hospital Compare Performance Measures and Mortality Rates." *Journal of the American Medical Association* 296 (22): 2694–702.
- Williams, R. L. 2000. "A Note on Robust Variance Estimation for Cluster-Correlated Data." *Biometrics* 56 (2): 645–6.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2: Selection of PHQID Reference Conditions.

Appendix SA3: Sensitivity Analysis for Heart Failure and Pneumonia Patients.

Appendix SA4: Cost and Outlier Figures.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.